

Available online at www.sciencedirect.com**ScienceDirect**

Procedia - Social and Behavioral Sciences 195 (2015) 1905 – 1914

Procedia
Social and Behavioral Sciences

World Conference on Technology, Innovation and Entrepreneurship

Solving The Yield Optimization Problem for Wafer to Wafer 3d Integration Process

Marwa Harzi ^{a*}, Hashem Abusenenh ^b, Saoussen Krichen ^a^a LARODEC, Institut Supérieur de Gestion, Tunisia^b Palestine Ahliya University College, Palestine

Abstract

Three dimensional integrated circuits (3D ICs) that stack multiple dies vertically using Through Silicon Vias (TSVs) have gained wide interests of the semiconductor industry. Fabricating these 3D ICs using wafer to wafer stacking has several advantages including: high throughput, high TSV density... However, one of the major challenges of the wafer to wafer stacking approach is the low compound yield. Various techniques have been presented in the literature to address this important problem. This paper investigates the compound yield improvement for wafer to wafer stacking method. To solve this problem, we propose two approaches: mathematical multi-dimensional axial assignment model using ILP and matching pre-tested wafers based Tabu search algorithm method. We focus on the performance of these two algorithms to solve the problem within an interesting running time and maximum yield. Finally, we present the results of our thorough computational study. The obtained results and the comparison with other algorithms show that our two algorithms gives better solutions.

© 2015 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Peer-review under responsibility of Istanbul Univeristy.

Keywords: 3D ICs, wafer to wafer, compound yield, assignment problem, wafer matching, Tabu search algorithm.

1. Introduction

The microelectronics industry is still evolving for over fifty years. A gain in manufacturing cost, a continuous miniaturization of electronic components while maintaining an increase in speed and performance of these are the major objectives that push this evolution (Roussel et al., 2008; Mukta and Subramanian, 2011). However, the

* Corresponding author.

E-mail address: harzimarwa@yahoo.fr

realization of these objectives attained physical limitations that call into question the planar approach (2D) used until now. Indeed, these limitations make appear parasitic effects such as increased interconnect delay, which could degrade the performance of electronic components. In addition, the current needs from consumers who use technology require diversification of functionalities (Gregory et al., 2009). A simple example can be found in our daily life which is the mobile phone. It allows taking pictures, making calls on Visio Mode while connecting to the Internet; Moreover, It permits the localization by using an integrated GPS system. The evolution of the planar approach (2D) used to this day to achieve these structures is very expensive and requires a complex design in order to group these different technologies on a same support (Roussel et al., 2008). It also results in an increase in the interconnection lines to connect each module and this will lead to increased delays in interconnections, reducing by the same occasion the performance of the final device.

By answering these problems met during the miniaturization and multi-functionality of devices that several actors from the world of microelectronics move to a new and an innovative concept which is three-dimensional (3D) integration (Chuan et al., 2008; Roussel et al., 2008; Mukta and Subramanian, 2011). This concept of 3D integration consists in stacking different types of substrates, then to electrically interconnecting them using a vertical connection crossing the silicon layers (TSV) (Taouil and Hamdioui, 2010; M. Taouil and Said, 2012). Several technological steps are to do in order to realize a 3D structure. It includes mainly the assembly of the substrates (or chips), the thinning of substrates once they are assembled and the realization of vertical connections through the silicon to permit the electrical connection (also called TSV for Through Silicon Vias) (Chuan et al., 2008). Various stacking methods can be used to assemble two substrates: wafer to wafer (WTW), die to wafer (DTW) and die to die (DTD) (Gregory et al., 2009; Taouil and Hamdioui, 2010; Mottaqiallah and Said, 2011; Roussel et al., 2010). These methods play an important role in determining the final yield of 3D ICs. The simplicity of the manufacturing process is the main advantage of WTW Stacking. However, without “Known good die” information, the yield suffers from a serious yield loss (Mottaqiallah and Said, 2011). On the other hand, DTW and DTD methods, requires a more complex manufacturing process that explore only good dies, thus resulting in higher yield compared to WTW stacking method. We show through the present study the powerfulness of the two proposed approach to solve the yield optimization problem for wafer to wafer 3D integration. Our problem was studied from an exact and an approximate point of view. The exact method called a multi-dimensional axial assignment model using ILP (MAA-ILP) through this one the problem is stated mathematically and solved using a commercial ILP solver named CPLEX solver. Then, the approximate method is a matching pre-tested wafers based tabu search algorithm (MW-TS). The proposed tool firstly inputs the basic parameters of the problem then by applying the TS algorithm the obtained result represent a set of wafer combinations that gives the yield in descending order. To check the validity of the proposed tool, we address a set of instances.

The reminder of this paper is structured as follows: In Section 2, we start with a brief overview of the related literature review. The yield improvement problem for wafer to wafer 3D integration process is described and stated with a mathematical example in section 3. Then, the main steps of the proposed resolution methods are also outlined in section 3. After that Section 4, offer a comprehensive set of experimental results. And finally, Section 5 presents conclusions drawn from this works.

2. Related Prior Work

3D integration technology has become an increasingly active research topic. In (Mukta and Subramanian, 2011), the authors have made a review about the 3D integration and the family of technologies (wafer to wafer, die to die and die to wafer) which enable the stacking of layers with vertical connections between them. In the same context, the advantages offered by 3D integration, the potential applications and the classification of 3D ICs have been discussed at the level of the chapter (Roussel et al, 2010). Since there is an increasing attention, many papers on the topic of yield optimization for 3D ICs have been published in the literature (Qiang et al., 2012). In particular, the yield loss in wafer to wafer stacking method has been studied by several authors (Taouil and Hamdioui, 2010), (Gregory et al., 2009; Smith et al, 2007; Mottaqiallah and Said, 2011; M. Taouil and Said, 2012; Vicent.B et al, 2013; Roussel et al, 2010). In order to solve the problem of yield loss especially in the wafer to wafer approach, several methods have been proposed. For example, the contributions (Mottaqiallah and Said, 2011 and Vicent.B et al, 2013) which are on the topic of yield improvement for 3D wafer to wafer stacked memories by using the layer redundancy method. In

the papers (Gregory et al., 2009 and Vicent.B et al, 2013), the problem is formulated as a multi-index assignment problem. We demonstrate the effectiveness of this method up to large numbers of wafer stacks. A survey on multi-dimensional assignment problem can be found in (Peter and William, 2012) and chapter 10 of (M.Taouil and Said, 2012). In other contributions, to address the yield loss in 3D ICs the authors have been proposed the matching technique (Taouil and Hamdioui, 2010; Gregory et al., 2009; Roussel et al, 2010). The results demonstrate that this technique can improve the yield (reaching up to 25% in (Gregory Smith et al., 2007) and 13.4% in (Taouil and Hamdioui, 2010)) relative to yield of random wafer to wafer stacking. The paper (Roussel et al, 2010), reports that the stack yield increases between 0.5% to 10%.

3. Problem Description

In the following Table 1, we enumerate the major symbols that will be used in this paper.

Table 1: Notations

Symbols	Description
W	Number of Wafers
L	Number of Wafers Lots
D	Number of dies per wafer
n	Number of Stacks
F_w	Number of Faulty dies per wafer
G_w	Number of Good dies per wafer

In this context, the problem addressed in this paper which is “the yield optimization problem in the semiconductor industry” can be informally described as follows: There are L lots of wafers named wafer lots, where each wafer lot contains exactly W wafers. A wafer consists of a string of bad dies and good dies; in our context this translates to “0” in case of bad die, and “1” in case of good die.

The objective is to form n stacks by integrating one wafer from each lot while maximizing the yield. Since, our studied problem is based mainly on the maximization of yield especially for wafer to wafer 3D integration process, at this level, we will detail how the yield of this integration will be calculated.

3.1 Example

- *Yield Per Wafer*

$$\text{Yield Per Wafer} = \text{number of good dies} = \text{total number of dies} [\%] \quad (3.1.1)$$

Let's consider the example in Figure 1 with $D = 16$, $F_w = 5$ and $G_w = 11$

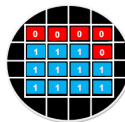


Figure 1: Example of Yield per wafer

By applying the rule (3.1.1), the yield per wafer = $G_w = 11 = 16 = 68$; 75%

- *Yield Wafers Stack*

$$\text{Yield Wafers Stack} = \text{Number of good stack} = \text{Total number of wafers stacks} [\%] \quad (3.1.2)$$

Let's consider the example in Figure 2 with $D_{w1} = D_{w2} = 6$, $G_{w1} = 3$ and $G_{w2} = 4$

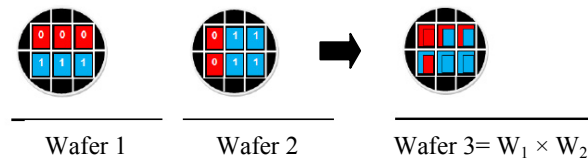


Fig 2: Example of Yield wafers stack

By applying the rule (3.1.2), the yield wafers stack ($W3$) = $GW3=D=2=6=33$; 33%

Since the yield optimization problem in wafer to wafer 3D integration is an NP-hard problem (Vicent. B et al., 2013), three approaches are typically employed to solve it: heuristics, approximation methods and exact methods. Only instances of small size can be solved to optimality using exact solution methods. While heuristics do not provide guarantees about the solution quality, they are useful in practical contexts due to their speed and ability to handle giant instances. That is why, we have used the MAA-ILP and the MW-TS as exact and approximate resolution method.

3.2 Method 1: A multi-dimensional axial assignment model using ILP

Due to the fact that in the instances of the yield optimization problem, the number of good dies in each wafer is usually much more than the number of bad dies, it make sense to minimize the number of bad stack dies in order to maximize the functional yield of wafer to wafer 3D integration.

In what follow Table 2, we give a straightforward mathematical formulation of the yield optimization problem in wafer to wafer 3D integration as a multi-dimensional axial assignment model using ILP (Frits and R. Spieksma, 2000).

Table 2: The mathematical model of the multi-dimensional axial assignment method using ILP

Parameters	
▪ $x_{i1, i2, iw}$: A binary variable that is true when wafer $i_1 \in \{1, \dots, N\}$, wafer $i_2 \in \{1, \dots, N\}$ and wafer $i_w \in \{1, \dots, N\}$ are stacked into a 3D integrated circuit.	
▪ $F_{i1, i2, iw}$: Denote the number of bad dies resulting from integrating the i_1, i_2, i_w wafers.	
Formulation	
Given W wafers each with D die, the functional yield maximization problem can be formulated as follows:	
Minimize $\sum_{i1=1}^n \dots \sum_{iW=1}^n Y_{f1, \dots, fW} \times x_{f1, \dots, fW}$	(3.2.1)
Subject to	
$\sum_{i1=1}^n \dots \sum_{iW=1}^n x_{f1, \dots, fW} = 1$; for $f1, \dots, fW = 1, \dots, n$	(3.2.2)
$x_{f1, \dots, fW} \in \{0, 1\}$; for $f1, \dots, fW = 1, \dots, n$	(3.2.3)
Objective Function	
Equation (3.2.1) allows to minimizing the number of bad die in order to maximize the functional yield of wafer to wafer 3D integration in accordance with the set of system constraints.	
System Constraint	
Constraint (3.2.2) ensures that each wafer in any lot participates in exactly one 3D wafer stack. And constraint	

(3.2.3) is a binary constraint.

We will solve our problem with a method called MAA-ILP, the resolution will be detailed as follow: A code written in C++ received as an input a file named “input.txt” that contains the problem instances. Then, this code will generate a file called “Solution.lp” which contains the objective function of our problem. The obtained file will be solved by “CPLEX” solver.

3.3 Method 2: A matching pre-tested wafers based Tabu Search algorithm

For the second method which is the MW-TS, in order to combine the advantages of both methods heuristic and meta-heuristic, we have used as a meta-heuristic the Tabu Search algorithm and like a heuristic the matching process. These two concepts will be detailed in the following sections.

- *Matching pre-tested wafers*

As above mentioned, wafer to wafer stacking approach suffers from a low compound yield. Wafer matching has been researched to deal with this drawback by many authors (Taouil and Hamdioui, 2010; Gregory et al., 2009; Mottaqiallah and Said, 2011, M. Taouil and Said, 2011; Roussel et al, 2010). In order to solve the yield maximization problem, the matching pre-tested wafers process finds the best n wafer stacks that maximize the total yield. This concept is a technique that exists to improve the yield by finding out the best wafer combinations that would result in higher yield, given that the wafers were tested before the stacking. So wafer matching is only possible if pre-bond test results are available for all dies.

Algorithm 1: Matching Heuristic

Input: L wafer Lots each with W wafers

for each stack in the N^K stacks: **do**

 Calculate the number of good 3D ICs resulting from the matching process.

end for

Output: W^L wafer stacks in descending order according to the yield stack they produce.

- *Tabu Search algorithm*

The TS approach operates in 3 main phases:

- 1) Start from an initial solution that can be randomly generated or heuristically designed.
- 2) Improve the current solution by applying one or more neighborhood search strategies.
- 3) Update the best recorded solution by comparing its fitness function with neighbor solutions.

Algorithm 2: Tabu Search algorithm

Require: L.lot = list of Lot

Require: L.wfr = list of wafers

Require: ListT = NULL

Find the initial solution S_0

$S^* \leftarrow S_0$

$Y^* \leftarrow f(S_0)$

while Time \leq Time_{max} **do**

$S = \text{exchange}(S^*)$ and S not in ListT

if $f(S) < Y^*$ **then**

$S^* \leftarrow S$

$f^* \leftarrow f(S)$

 Update ListT

end if

end while

1. Encoding and initialization

The encoding of a solution is designed in such a way to maximize the yield of wafer to wafer 3D integration. In the 3D stacking, from each lot of wafers one wafer is taken to be stacked. As shown in Figure 3, our solution is encoded in the form of a vector that contains all treated wafers Lots

Each cell of the vector is composed of the entire wafers for each lot. If the wafer takes the value “1” so this one is selected to be used in the process of stacking for the 3D ICs, else if “0” it is not concerned.

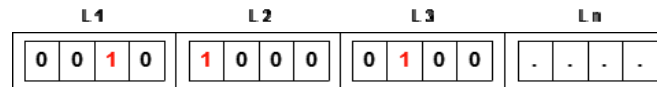


Fig 3: Solution Encoding

2. Neighborhood generation

Given a current solution, the exchange local search techniques are used sequentially to alternatively generate the neighborhood. Our incentive behind using exchange techniques is to diversify the search and increase the probability of identifying the optimal solution.

Exchange Technique: This technique swaps two wafers belonging to two different lots.

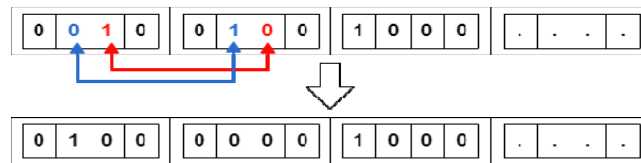


Fig 4: Exchange Technique

As shown in Figure 4, wafer 2 from lot 1 and wafer 2 from lot 2, are switched with wafer 3 from lot 1 and wafer 3 from lots 2 respectively. This yields to a new solution that gives more of yield.

3. Fitness function

The evaluation of solutions is one of the fundamental procedures in any optimization approach and that's the case too in improvement yield in wafer to wafer 3D integration process. Each solution is evaluated according to its fitness that corresponds, for our problem, the objective function value as all solutions are feasible. Throughout the search process, the solution having best fitness is recorded and updated while iterating in the Tabu Search algorithm. Hence, the fitness function of each currently evaluated solution S is computed as follows:

$$Fitness(w) = f(w) = \sum_{i=1}^n \dots \sum_{i_k=1}^n Y_{i_1 i_2 \dots i_k} \times X_{i_1 i_2 \dots i_k}$$

Our fitness function is to maximize the yield of wafer to wafer 3D integration. Where x represents the solution already encoded in the Tabu Search.

4. Stopping rule

After a predefined time “Time_{max}”, experimentally set depending on the problem size, the TS algorithm is over.

- *Main resolution steps*

Step 1: Data inputs

A first step Figure 5 consists in inputting the basic parameters of the problem as: Number of lots $L = 3$, Number of wafers per lot $W = 3$, Number of dies per wafer $D = 10$.

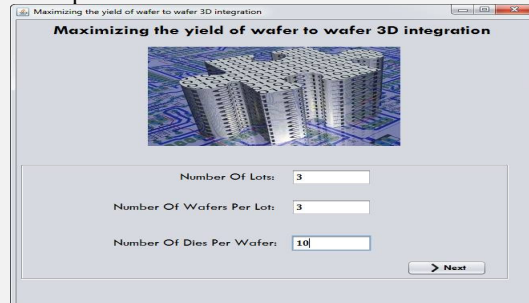


Figure 5: The first screenshot “Data inputs”

Step 2: Generating values of dies

Once the problem size is set dies are to be selected from an existing database or entered as new data Figure 6. The generating values of dies is a succession of “0” and “1”, where “1” represents the good dies while “0” represents the bad dies.

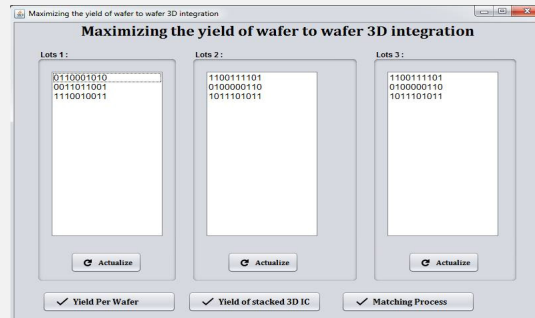


Figure 6: The second screenshot “Generating values of dies”

Step 3: Matching process

As a result, we have a set of matching wafer stack. This set contains W^L wafer stacks in descending order according to the number of good die they produce, in other words according to the yield of each stack.

In this example Figure 7, we have $W = 3$ and $L = 3$ hence the number of matching wafer stacks is equal to $3 \times 3 = 9$ stacks.

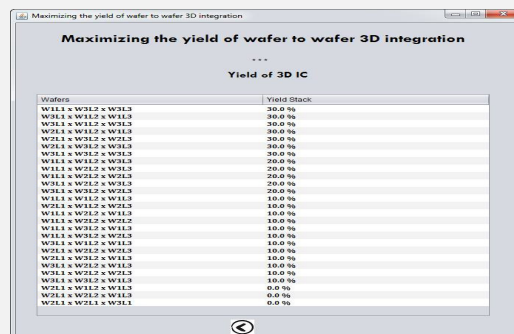


Figure 7: The third screenshot “Matching process”

Table 3: The main resolution steps

5. Experimental results and main remarks

In order to improve the effectiveness and the good results of our two proposed approaches in maximizing the functional yield of wafer to wafer 3D integration, we exposed a set of comprehensive experiments to clarify and analyze them.

3.4 Description of the instances

In this section, we will describe the format of the instances on which we will work along the testing process: The format of the instances is as follows:

“Number of Lots - number of Wafers per Lot - number of Dies per Wafer - generation of dies values”.

In our case, an instance is denoted by L-W-D with:

L = 3	W \in {10, 20, 50}	D \in {20, 50}
-------	----------------------	------------------

The generating values of dies is a collection from “0” and “1”, where “1” represents the good dies while “0” represents the bad dies.

3.5 The effect of yield per wafer on the yield stack

Firstly, we investigate the impact of the defect density (ie. the density of default dies) per wafer on the final yield of the produced 3D ICs. We compare the performance of the two proposed integration methods **MAA-ILP** and **MW-TS** at different defect densities. Concerning the other two methods, **OP-ILP** and **IMH** represents methods used by (Gregory Smith et al., 2009). We have put the results of these two methods, in order to compare them with the results of our proposed methods. (-) this means that is not indicated in the article of (Gregory Smith et al., 2009).

We set the number of wafers in the 3D stack to be equal to $W = 3$ and vary the defect density to result in yields (Yw) from 50% to 99% per wafer. In table, we report values of the overall yield of 3D ICs (Ys) for each integration algorithm.

The results show that the two proposed integration algorithms giving as a result an overall improvement yield compared to the random integration yield and also to the OP-ILP and IMH methods yield. As illustrated in Table 4, whenever the number of bad dies decreases the yield per wafer (Yw) increases and consequently the yield stack (Ys) also increases.

Table 4: The Impact of defect density per wafer on the yield

Methods	Yield per wafer					
	50%	60%	70%	80%	90%	99%
Random	12.5%	21.6%	34.3%	51.2%	72.9%	97.02%
MAA-ILP	19.95%	29.5%	41.9%	58.07%	77.97%	98.13%
MW-TS	21.69%	31.9%	44.67%	60.38%	79.96%	99.53%
OP-ILP	15.08%	-	37.29%	-	74.41%	-
IMH	14.75%	-	36.95%	-	74.41%	-

3.6 The effect of the number of wafers per wafer lot on running time stack

We propose to test the impact of the number of wafers per lot on the running time stack (ts) which is expressed in seconds (s). For a definite number of dies, the number of wafers will be changed each time to determine their impact.

That is why, we have three case, the first one with a number of wafers per wafer lot $W = 20$, $W = 50$ and $W = 50$ in each case we change the number of wafers per lot. We have noticed that the two proposed algorithms give good and effective results shown in Table 5. But we have seen that the MAA-ILP approach gives better results than the MW-TS approach. Also, the running time stack (ts) increases for each time that the numbers of dies per wafer increase.

Table 5: The impact of the number of wafers per wafer lot on execution time stack

	Methods	Yield per wafer					
		50%	60%	70%	80%	90%	99%
W=10	MAA-ILP	0,34	0,05	0,034	0,02	0,012	0,001
	MW-TS	0,501	0,123	0,077	0,07	0,077	0,05
W=20	MAA-ILP	1,986	0,876	0,21	0,19	0,08	0,09
	MW-TS	2,599	1,098	0,315	0,331	0,302	0,271
W=50	MAA-ILP	6,023	5,564	4,64	4,13	3,65	2,014
	MW-TS	7,554	6,94	5,99	5,399	4,601	3,043

3.7 The effect of the number of dies per wafer on running time stack

In this important experiment, we study the impact of the number of dies per wafer on running time stack (ts) which is also expressed in seconds (s).

At this level, in each case we fixed the number of wafers and we vary the number of dies per wafer. We have two case, the first one with a number of dies $D = 20$ and the second one $D = 50$, and in each case we change the number of wafers per lot. In Table 6 we note that, in each time if the yield increases the running time stack decrease. Also, the method MAA-ILP gives execution time stack less than the MW-TS method.

Table 6: The impact of the number of dies per wafer on running time stack

	Methods	Yield per wafer					
		50%	60%	70%	80%	90%	99%
D=20	MAA-ILP	0,23	0,09	0,021	0,012	0,01	0,003
	MW-TS	0,501	0,123	0,077	0,07	0,077	0,05
D=50	MAA-ILP	1,765	1,34	1,02	0,91	0,875	0,532
	MW-TS	2,783	2,14	1,6	1,301	1,199	1

6. Conclusion

In this paper, we have studied an important combinatorial optimization problem named “the yield improvement in wafer to wafer 3D integration”. In the above work, we have proposed two techniques to solve the problem: the first one called “A multi-dimensional axial assignment model using ILP” which guarantees the optimality of the results, but at a finite number of wafers and dies this approach failed because the exact methods supports only small instances that is why we used a second methods named “Matching pre-tested wafers based Tabu Search algorithm” this one allow to find good solutions but not in any way guarantee the optimality of these. The computational experiment shows that the two suggested approaches MAA-ILP and MW-TS are effective in terms of CPU runtime and meeting the objective of yield improvement problem in wafer to wafer 3D integration by generating a maximized yield stack.

References

- Brandon Noia, Krishnendu Chakrabarty, Erik Jan Marinissen. (2012). Optimization Methods for Post-Bond Testing of 3D Stacked ICs, Journal of Electronic Testing, 28, 103-120.
- Chuan Seng Tan, Ronald J. Gutmann, L. Rafael Reif. (2008). *Wafer Level 3-D ICs Process Technology*, chapter *Overview of Wafer-Level 3D ICs*, 1-11.
- Frits. C & R. Spijksma. (2000). *Nonlinear Assignment Problems*, chapter *Multi Index Assignment Problems: Complexity, Approximation, Applications*, 1-12.

- Gregory Smith, Sherief Reda, Larry Smith. (2009). Maximizing the functional yield of wafer-to-wafer 3-d integration. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 17, 1357–1362.
- Mukta G. Farooq & Subramanian S. Iyer (2011). 3D integration review, *Science China Information Sciences*, 54, 1012-1025.
- Mottaqiallah Taouil & Said Hamdioui. (2011) Layer redundancy based yield improvement for 3d wafer-to-wafer stacked memories. 16th IEEE European Test Symposium (ETS), Trondheim, 45 – 50
- M. Taouil & Said Hamdioui. (2012). Yield improvement for 3d wafer-to-wafer stacked memories. *Springer J Electron Test*, 523–534.
- Peter M.Hahn & William L.H. (2010). Lower bounds for the axial three-index assignment problem. *European Journal of Operational Research*, 13, 654–668.
- Qiang Xu, Li Jiang, Huiyun Li, Eklow. B. (2012) Yield enhancement for 3D-stacked ICs: Recent advances and challenges, 17th Asia and South Pacific - Design Automation Conference (ASP-DAC), Sydney NSW.
- Roussel. P, Velenis. D, Verbree. J, Marinissen. E.J. (2010). On the cost-effectiveness of matching repositories of pre-tested wafers for wafer-to-wafer 3d chip stacking. 15 th IEEE European Test Symposium (ETS), Praha.
- Smith. G, Smith.Larry, Hosali. S. (2007). Yield considerations in the choice of 3d technology. International Symposium on Semiconductor Manufacturing (ISSM).
- Taouil .M, Hamdioui .S, Verbee .J, Marinissen .E.J. (2010) On maximizing the compound yield for 3D Wafer-to-Wafer stacked ICs, IEEE International Test Conference (ITC), Austin TX.
- Vincent.B, Rodolphe.G, Trivikram.D, Marin.B, Frits C.R.S. (2013). Approximation algorithms for the wafer to wafer integration problem. *Springer -Verlag Berlin Heidelberg*, 7846, 286–297.